Docket No.: M4065.0656/P656

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICATION FOR U.S. LETTERS PATENT

Title:

NON-VOLATILE MEMORY STRUCTURE

Inventor:

Terry L. Gilton

DICKSTEIN SHAPIRO MORIN &
OSHINSKY LLP
2101 L Street NW
Washington, DC 20037-1526
(202) 828-2232

NON-VOLATILE MEMORY STRUCTURE

FIELD OF THE INVENTION

[0001] The present invention relates to the field of semiconductor memory devices and, more particularly, to a flash memory device.

BACKGROUND OF THE INVENTION

[0002] A nonvolatile memory is a type of memory that retains stored data when power is removed. There are various types of nonvolatile memories including e.g., read only memories (ROMs), erasable programmable read only memories (EPROMs), and electrically erasable programmable read only memories (EEPROMs), and flash memory.

[0003] Flash memory is often used where regular access to the data stored in the memory device is desired, but where such data is seldom changed. For example, computers often use flash memory to store firmware (e.g., a personal computer's BIOS). Peripheral devices such as printers may store fonts and forms on flash memory. Wireless communications devices such as cellular and other wireless telephones use flash memory to store data and their operating systems. Portable electronics such as digital cameras, audio recorders, personal digital assistants (PDAs), and test equipment use flash memory cards as a storage medium.

[0004] Flash memory cells make use of a floating-gate covered with an insulating layer. There is also a control gate which overlays the insulating layer. Below the floating gate is another insulating layer sandwiched between the floating gate and the cell substrate. This insulating layer is an oxide layer and is often referred to as the tunnel oxide. The substrate contains doped source and drain regions, with a channel region disposed between the source and drain regions. The floating-gate transistors generally include n-channel floating-gate field-effect transistors, but may also include p-channel floating-gate field-effect transistors. Access operations are carried out by applying biases to the transistor.

[0005] In a flash memory device, cells are often organized into blocks and the charge state of the floating gate indicates the logical state of the cell. For example, a charged floating gate may represent a logical "1" while a non-charged floating gate may represent a logical "0." A flash memory cell may be programmed to a desired state by first erasing the cell to a logical "0" and, if necessary, writing the cell to a logical "1." Typically, flash memory devices are organized so that a write operation can target a specific cell while an erase operation affects an entire block of cells. Changing any portion of one block therefore requires erasing the entire block and writing those bits in the block which correspond to a logical "1".

[0006] Referring now to Fig. 1, a conventional flash memory cell 10 includes a source region 26 and a drain region 28. The source 26 and drain 28 have an N+ type conductivity formed in a P-type substrate 20. The memory cell 10 has a stack-gate configuration which includes a cap layer 22 formed over a control gate 18 formed over an insulating layer 16 formed over a floating gate 14 formed over a tunnel oxide layer 12. The floating gate 14 is formed of a first polysilicon layer and the control gate 18 is formed of a second polysilicon layer. The floating gate 14 is isolated from the control gate 18 by the insulating layer 16 and from a channel region 30 of the substrate 20 by the tunnel oxide layer 12. The tunnel oxide layer is generally about 100 Angstroms thick.

[0007] Referring now to Fig. 2, the conventional flash memory cell 10 is shown during a programming operation. A positive programming voltage Vp of, e.g., about 12 volts is applied to the control gate 18. The positive programming voltage Vp attracts electrons 32 from the P-type substrate 20 and causes them to accumulate at the surface of channel region 30. A voltage on drain 28 Vd is increased to, e.g., about 6 volts, and the source 26 is connected to ground Vs. As the drain-to-source voltage increases, electrons 32 flow from the source 26 to the drain 28 via channel region 30. As electrons 32 travel toward the drain 28, they acquire substantial high kinetic energy and are typically referred to as "hot" electrons.

[0008] The voltages at the control gate 18 and the drain 28 create an electric field in the oxide layer 12, which attracts the hot electrons and accelerates them toward the floating gate 14. At this point, the floating gate 14 begins to trap and accumulate the hot electrons. This is a charging process. As the charge on the floating gate 14 increases, the electric field in the oxide layer 12 decreases gradually and eventually loses its capability of attracting more hot electrons to the floating gate 14. At this point, the floating gate 14 is fully charged. The cell 10 will turn on when the voltage on the control gate 18 is brought to the threshold voltage level of the cell 10. Sense amplifiers are used in the memory to detect and amplify the state of the memory cell during a read operation.

[0009] Electrons are removed from the floating gate 14 to erase the memory cell. Fowler-Nordheim (FN) tunneling may be used to erase the memory cell 10. The erase procedure is accomplished by electrically floating the drain 28, grounding the source 26, and applying a high negative voltage (e.g., -12 volts) to the control gate 18. This creates an electric field across the tunnel oxide layer 12 and forces electrons off of the floating gate 14 and to then tunnel through the tunnel oxide layer 12 back to the substrate 20.

[0010] The erase operation requires high voltages and is relatively slow. The high erase voltages are a fundamental problem arising from the high electron affinity of bulk silicon or large grain polysilicon particles used as the floating gate. This creates a very high tunneling barrier. Even with high negative voltages applied to the gate, a large tunneling distance is experienced with a very low tunneling probability for electrons attempting to leave the floating gate. This results in long erase times since the net flux of electrons leaving the gate is low. Thus, the tunneling current discharging the gate is low. In addition, other phenomena result as a consequence of this very high negative voltage. Hole injection into the oxide is experienced which can result in erratic over erase, damage to the gate oxide itself, and the introduction of trapping states. Accordingly, there is a desire and need for a new flash memory cell architecture, which overcomes the aforementioned problems.

SUMMARY OF THE INVENTION

[0011] The present invention is directed to a flash memory cell in which the floating gate is implemented using a programmable conductance random access memory structure instead of the traditional polysilicon layer. Instead of storing or removing electrons from a polysilicon layer, the programmable conductance is switched between its low and high resistive states to operate the cell. The resulting cell can be erased faster and has better endurance (i.e., can withstand a greater number of erase/write cycles) than a conventional flash memory.

BRIEF DESCRIPTION OF THE DRAWINGS

[0012] The foregoing and other advantages and features of the invention will become more apparent from the detailed description of exemplary embodiments of the invention given below with reference to the accompanying drawings in which:

[0013] Fig. 1 is an illustration of a prior art flash memory cell;

[0014] Fig. 2 is an illustration of how the prior art flash memory cell of Fig. 1 is programmed;

[0015] Fig. 3 is an illustration of an embodiment of a flash memory cell of the invention;

[0016] Figs. 4-5 are cross-section illustrations of a flash memory cell of the invention at various stages of fabrication;

[0017] Fig. 6 is an illustration of the flash memory cell of the invention coupled to a word line and a bit line;

[0018] Fig. 7 is an illustration of how the flash memory cell of an embodiment of the invention may be programmed;

[0019] Fig. 8 is an illustration of a flash memory device incorporating the flash memory cell of an embodiment of the invention; and

[0020] Fig. 9 is an illustration of a computer system incorporating the flash memory device of Fig. 8.

DETAILED DESCRIPTION OF THE INVENTION

[0021] In the following detailed description, reference is made to the accompanying drawings, which form a part hereof, and in which is shown by way, of illustration of specific embodiments in which the invention may be practiced. It is to be understood that other embodiments may be utilized, and that structural, logical and electrical changes may be made without departing from the spirit and scope of the present invention.

[0022] The terms "wafer" and "substrate" are to be understood as including silicon, silicon-on-insulator (SOI) or silicon-on-sapphire (SOS) technology, doped and undoped semiconductors, epitaxial layers of silicon supported by a base semiconductor foundation, and other semiconductor structures. Furthermore, when reference is made to a "wafer" or "substrate" in the following description, previous process steps may have been utilized to form regions or junctions in the base semiconductor structure or foundation. In addition, the semiconductor need not be silicon-based, but could be based on silicon-germanium, germanium, or gallium-arsenide.

[0023] Referring now to Fig. 3, there is shown a flash memory cell 100 of an embodiment of the invention. The cell 100 uses a programmable conductance structure 115, 117, 119 to replace the floating gate of the conventional flash memory cell (Figs. 1-2). Accordingly, a non-polysilicon variable resistance material capable of being switched from a first logic state to a second logic state is provided to promote electron tunneling through a tunnel oxide layer. Unlike a floating gate structure, electrons are not stored to operate the flash memory cell; instead, the programmable conductance structure is switched between resistance states. This results in faster erase times and greater endurance for the cell 100.

[0024] The programmable conductance structure 115, 117, 119 includes a programmable conductance in the form of a metal-doped chalcogenide glass disposed between two electrodes (conductive layers). The primary advantage of using a programmable conductance is that programmable conductance structures are not susceptible to the leakage current induced damage experienced by floating gates during erase cycles. Thus, a flash memory device utilizing a variable resistance material including a programmable conductor will have a greater endurance, as measured by the number of times each cell can be rewritten (i.e., erased and written).

[0025] Figs. 4 and 5 are cross-sectional views of a flash memory cell 100 at various stages of fabrication in accordance with the invention. Referring to Fig. 4, a gate dielectric layer, often referred to as gate oxide layer or a tunnel oxide layer (hereinafter T.O. layer) 112, is formed over a substrate 110, such as a wafer of single crystalline silicon (Si) or other material. The substrate 110 may be implanted with a p-type dopant to produce a p-type substrate.

[0026] The tunnel oxide layer 112 comprises a dielectric material, which preferably comprises an oxide material. The oxide may be formed by thermal or other oxidation

techniques. Other dielectric materials may be used for the T.O. layer 112. Specific examples include silicon oxides, silicon nitrides and silicon oxynitrides.

[0027] The structure 114 of the flash memory cell 100 includes a first electrode 115, a variable resistance material 117, and a second electrode 119 which are formed over the tunnel oxide layer 112. An insulating cap layer 122 is generally formed overlying the second electrode 119.

[0028] Although layer 117 is shown as a single material, it should be recognized that layer 117 may itself be comprised of a plurality of layers. For example, in one exemplary embodiment, layer 117 is comprised of a first layer comprising $Ge_{40}Se_{60}$, a second layer (formed over the first layer) comprised of Ag, and/or Se (e.g., the second layer might be a single layer of Ag_2Se, or a layer of Ag_2Se formed over a layer of Ag), a third layer formed over the second layer, the third layer comprising $Ge_{40}Se_{60}$, a fourth layer formed over said third layer comprising Ag, and a fifth layer formed over said fourth layer comprising $Ge_{40}Se_{60}$. In one exemplary embodiment, the first and third layer may be about 150 angstrom in thickness, and the fifth layer may be about 100 angstrom in thickness. The fourth layer may be approximately 200 angstrom in thickness. Finally, the second layer may be a 470 angstrom layer of Ag_2Se formed over a 35-50 angstrom layer of Ag.

[0029] The first electrode 115 is formed over the tunnel oxide layer 112. The first electrode 115 may comprise any conductive material, for example, tungsten, nickel, tantalum, aluminum, platinum, or silver, among many others.

[0030] The variable resistance material 117 is formed over the first electrode 115. One preferred material 117 comprises a chalcogenide glass. A specific example is germanium-selenide (Ge_xSe_{100-x}) containing a silver (Ag) component. A preferred germanium-selenide stoichiometric range of the resistance variable material 117 is between about $Ge_{18}Se_{82}$ to about $Ge_{43}Se_{57}$ and is more preferably about $Ge_{20}Se_{80}$.

[0031] One method of providing silver to germanium-selenide composition is to initially form a germanium-selenide glass and then deposit a thin layer of silver upon the glass, for example by sputtering, physical vapor deposition, or other known techniques in the art. The layer of silver is irradiated, preferably with electromagnetic energy at a wavelength less than 600 nanometers, so that the energy passes through the silver and to the silver/glass interface, to break a chalcogenide bond of the chalcogenide material such that the glass is doped or photodoped with silver. Another method for providing silver to the glass is to provide a layer of silver-selenide on a germanium-selenide glass.

[0032] The variable resistance material 117 is generally formed of dielectric material having a conductive material, such as silver, incorporated therein. The resistance of the variable resistance material 117 can be programmed between two bi-stable states having high and low resistances. The variable resistance material 117 is normally in a high resistance state. A write operation placing the material 117 into a low resistance state is performed by applying a voltage potential across the two electrodes 115, 119. A write operation placing the material into a high resistance state is performed by applying a reversed voltage potential across the two electrodes 115, 119. Accordingly the state of the flash memory cell will be determined by the potential applied to the structure 114.

[0033] The second conductive electrode 119 is formed over the variable resistance material 117. The second electrode 119 may comprise any electrically conductive material, for example, tungsten, tantalum, titanium, or silver, among many others. Typically, the second electrode 119 comprises silver.

[0034] A cap layer 122 is generally formed overlying the structure 114, and in particular, overlying the second electrode 119, to act as an insulator and barrier layer. The cap layer 122 contains an insulator and may include such insulators as silicon oxide, silicon nitride, and silicon oxynitrides. Preferably, the cap layer 122 is a silicon nitride, formed by

such methods as chemical vapor deposition (CVD). An example of the resulting structure is depicted in Fig. 3.

[0035] In Fig. 5, the tunnel oxide layer 112, the first electrode 115, the variable resistance material 117, second electrode 119, and the cap layer 122 illustrated in Fig. 4 are patterned to define the gate 150. It is noted that additional layers may form the gate 150, such as barrier layers to inhibit diffusion between opposing layers of adhesion layers to promote adhesion between opposing layers.

[0036] A source region 126 and a drain region 128 are formed adjacent to the gate 150 as conductive regions having a second conductivity type different than the conductivity type of the substrate 110. For example, the source and drain regions 126 and 128 are n-type regions formed by implantation and/or diffusion or n-type dopants, such as arsenic or phosphorus. The edges of the source and drain regions 126 and 128 are generally made to coincide with, or underlap, the gate edges. As an example, the source and drain regions 126 and 128 may be formed using light doped regions, as known in the art. Referring to Fig. 6, at least the second electrode 119 of the gate 150 is coupled to a word line 190. The source 126 and drain 128 are coupled to respective bit lines 192.

[0037] The sidewalls of the gate 150 are insulated using sidewall spacers 124 as shown in Fig. 3. The sidewall spacers 124 contain an insulator and may include the same materials as the cap layer 122. The sidewall spacers 124 are typically formed by blanket deposition an insulating layer, such as a layer of silicon nitride, over the entire structure and then anisotropically etching the insulating layer to preferentially remove the horizontal regions and to leave only the vertical regions adjacent the sidewalls of the gate 150.

[0038] Referring now to Fig. 7, to write (i.e., program) the memory cell 100, a positive programming voltage Vp of about 8 volts to 12 volts is applied to the second electrode 119. This positive programming voltage Vp attracts electrons 132 from p-type substrate 110 and causes them to accumulate toward the surface of channel region 130. A

drain 128 voltage Vd is increased to about 6 volts, and the source 126 is connected to ground. As the drain-to-source voltage increase, electrons 132 begin to flow from the source 126 to the drain 128 via channel region 130. The electrons 132 acquire substantially large kinetic energy and are referred to as hot electrons.

[0039] The voltage difference between the second electrode 119 and drain 128 creates an electric field through the tunnel oxide layer 112, this electric field attracts the hot electrons 132 and accelerates them towards the first electrode 115. The first electrode 115 starts to trap and accumulate the hot electrons 132, beginning the charging process. As the charge on the first electrode 115 increase, the electric field through tunnel oxide layer 112 decrease and eventually loses it capability of attracting any more of the hot electrons. At this point, the first electrode 115 has sufficient charge, such that the voltage potential across the two electrodes causes a conductive path to form across the variable resistance material 117 from the second electrode 119 to the first electrode 115. A threshold voltage Vt of the memory cells is equivalent to the voltage potential across the two electrodes which causes the conductive path to form. Since the typical non-programmed state of a variable resistance material 117 is the high resistance state (e.g., logical "0"), the memory element is programmed by an applied voltage to place the memory element into a low resistance state (e.g., logical "1"). The resistance between the two electrodes 115, 119 of the cell thus becomes a function of the presence or absence of a conductive path in the cell 100.

[0040] The memory cell 100 may be erased using Fowler-Nordheim (FN) tunneling. More specifically, the drain 128 is electrically floated, the source 126 is grounded, and a high negative voltage (e.g., about -12 volts) is applied to the second electrode 119. This creates an electric field across the tunnel oxide layer 112 and forces electrons 132 off of the first electrode 115 which then tunnel through the tunnel oxide layer 112 to the source region 126. Additionally, the conductive path begins to retract, which in turn increases the

resistance of the memory cell 100 to coincide with the high resistance state (e.g., logical "0").

[0041] A read operation is performed by sensing a difference caused by the memory cell 100 being in a first programmed state (e.g., logical "1") or a second programmed state (e.g., logical "0"). Referring now to Fig. 7, a read operation can be started by applying a reading voltage to the second electrode 119. The reading voltage is chosen so that when the memory cell 100 is in the first programmed state, an inversion layer 140 is formed in the channel region 130 below the tunnel oxide layer 112, and when the memory cell 100 is in the second programmed state, no inversion layer 140 is formed. The inversion layer 140 can be thought of as an extension of the source/drain regions 126/128 into the channel region 130. As discussed below, the presence or absence of an inversion layer 140 can be used to cause a difference in bit line capacitance or current flow through the cell 100.

[0042] A memory cell 100 in the first programmed state (and thus having an inversion layer 140) would have a greater bit line capacitance than the same memory cell 100 in the second programmed state (not having an inversion layer 140). The increase in capacitance in the first programmed state is due to the inversion layer 140 coupling an additional source/drain junction in the memory cell 100. Similarly, if a small (e.g., about 0.3-0.8 volt) forward bias is applied to the target bit line, a memory cell 100 in the first programmed state (and having an inversion layer) would have a higher level of forward current flow through the bit line than the same memory cell 100 in the second programmed state (and not having an inversion layer). The increase in current flow in the first programmed state is due to the inversion layer 140 creating a larger a larger effective diode area for the bit line forward current through the memory cell 100.

[0043] The memory cell 100 can then be read by conventionally sensing the above described differences between the two programmed states. For example, a sensing scheme may include precharging a target bit line and a reference bit line to respective reference

levels, coupling the target bit line to the memory cell 100 while the reading voltage is on, and sensing a difference in voltage between said target bit line and said reference bit line after a predetermined time.

[0044] Fig. 8 is an illustration of flash memory device 200. The flash memory device 200 includes a plurality of individually erasable blocks 201. Each block 201 includes a plurality of flash memory cells 100 (Fig. 3). The blocks 201 are coupled to a row control circuit 202 and a column control circuit 203, for addressing and controlling reading, writing, and erasing of one or more memory cells 100 (Fig. 3) of a selected block 201. The column control circuit 203 is also coupled to a write buffer 204, which holds data to be written and to input/output buffers 205 for buffering off-device communications. A controller 206, coupled to the row control circuit 202, column control circuit 203, and input/output buffers 205, coordinates the reading, writing, and erasing of the device 200.

[0045] Fig. 9 illustrates an exemplary processing system 900 which may utilize the memory device 200 of the present invention. The memory device 200 may be found, for example, in a memory component 908 of the system 900. The processing system 900 includes one or more processors 901 coupled to a local bus 904. A memory controller 902 and a primary bus bridge 903 are also coupled the local bus 904. The processing system 900 may include multiple memory controllers 902 and/or multiple primary bus bridges 903. The memory controller 902 and the primary bus bridge 903 may be integrated as a single device 906.

[0046] The memory controller 902 is also coupled to one or more memory buses 907. Each memory bus accepts memory components 908 which include at least one memory device 200 of the present invention. The memory components 908 may be a memory card or a memory module. Examples of memory modules include single inline memory modules (SIMMs) and dual inline memory modules (DIMMs). The memory components 908 may include one or more additional devices 909. For example, in a

SIMM or DIMM, the additional device 909 might be a configuration memory, such as a serial presence detect (SPD) memory. The memory controller 902 may also be coupled to a cache memory 905. The cache memory 905 may be the only cache memory in the processing system. Alternatively, other devices, for example, processors 901 may also include cache memories, which may form a cache hierarchy with cache memory 905. If the processing system 900 include peripherals or controllers which are bus masters or which support direct memory access (DMA), the memory controller 902 may implement a cache coherency protocol. If the memory controller 902 is coupled to a plurality of memory buses 907, each memory bus 907 may be operated in parallel, or different address ranges may be mapped to different memory buses 907.

[0047] The primary bus bridge 903 is coupled to at least one peripheral bus 910. Various devices, such as peripherals or additional bus bridges may be coupled to the peripheral bus 910. These devices may include a storage controller 911, a miscellaneous I/O device 914, a secondary bus bridge 915 communicating with a secondary bus 916, a multimedia processor 918, and a legacy device interface 920. The primary bus bridge 903 may also coupled to one or more special purpose high speed ports 922. In a personal computer, for example, the special purpose port might be the Accelerated Graphics Port (AGP), used to couple a high performance video card to the processing system 900.

[0048] The storage controller 911 couples one or more storage devices 913, via a storage bus 912, to the peripheral bus 910. For example, the storage controller 911 may be a SCSI controller and storage devices 913 may be SCSI discs. The I/O device 914 may be any sort of peripheral. For example, the I/O device 914 may be an local area network interface, such as an Ethernet card. The secondary bus bridge 915 may be used to interface additional devices via another bus to the processing system. For example, the secondary bus bridge may be an universal serial port (USB) controller used to couple USB devices 917 via to the processing system 900. The multimedia processor 918 may be a sound card, a video capture card, or any other type of media interface, which may also be coupled to

one additional devices such as speakers 919. The legacy device interface 920 is used to couple at least one legacy device 921, for example, older styled keyboards and mice, to the processing system 900.

[0049] The processing system 900 illustrated in Fig. 9 is only an exemplary processing system with which the invention may be used. While Fig. 9 illustrates a processing architecture especially suitable for a general purpose computer, such as a personal computer or a workstation, it should be recognized that well known modifications can be made to configure the processing system 900 to become more suitable for use in a variety of applications. For example, many electronic devices which require processing may be implemented using a simpler architecture which relies on a CPU 901 coupled to memory components 908 and/or memory devices 100.

[0050] While the invention has been described in detail in connection with the exemplary embodiment, it should be understood that the invention is not limited to the above disclosed embodiment. Rather, the invention can be modified to incorporate any number of variations, alternations, substitutions, or equivalent arrangements not heretofore described, but which are commensurate with the spirit and scope of the invention. Accordingly, the invention is not limited by the foregoing description or drawings, but is only limited by the scope of the appended claims.